

YOR920000324US1

12

Clean Copy of Amended Title

AUTOMATED SET UP OF WEB-BASED
NATURAL LANGUAGE INTERFACE

Clean Copy of Amended Paragraphs

Paragraph on page 1, lines 14 to 16, now reads as follows:

a1

The present invention generally relates to natural language systems and, more particularly, to an automated method for setting up a Web-based natural language interface.

Paragraph beginning on page 1, line 18, and continuing to page 2, line 2, now reads as follows:

a2

The World Wide Web (WWW) portion of the Internet has seen an explosion of Web sites for various individual and business purposes. This in turn has led to a growing industry in Do It Yourself (DIY) software and Web design services to assist those who want to set up a Web site.

Paragraph on page 2, lines 11 to 13, now reads as follows:

a3

It is therefore an object of the present invention to provide a procedure that automates the process of setting up an instance of a natural language interface for a Web site.

Paragraph on page 2, lines 16 to 22, now reads as follows:

a4

This invention, by automating the process of setting up a new Web site, enables a new interface to be created by anyone. Subsequent manual tuning of the interface is possible and much easier to do than creating an interface from scratch. The invention solves the problem by bringing together a number of ideas and techniques, some of which have been used in natural language processing for other purposes. In order to set up an instance of a natural language interface, it is

Q4

necessary to

Paragraph beginning on page 4, line 16, and continuing to page 5, line 17,
now reads as follows:

Q5

Referring now to the drawings, and more particularly to Figure 1, there is shown a flow diagram of the automated set up procedure. A program implementing a Web crawler is invoked in function block 11, beginning at the home page of the site for which a natural language interface is to be generated. The output of this module is a file of Web pages in HyperText Markup Language (HTML). In function block 12, the Uniform Resource Locators (URLs) of the Web pages are processed to induce a hierarchy of topics for the site and the HTML formatted pages are converted to the appropriate standard format. In a preferred implementation of the invention, the standard format is eXtensible Markup Language (XML). In function block 13, sparse n-grams are extracted from each page to serve as index terms for the page. The index terms are used to set up an answer generator (search engine) for the page in function block 14. In function block 15, a set of sparse n-grams is generated for each of the topics found in function block 12 by grouping together all the documents having that topic. Those n-grams satisfying some criterion for significant association with the topic are saved. In a preferred implementation of the invention, the criterion used is the chi-square measure. Optionally, another statistical test can be made to associate a confidence measure with each rule. In the preferred implementation of the invention, the confidence measure is the percentage of time the underlying n-gram occurs in the topic. Each sparse n-gram is converted to a rule in which each term of an n-gram is a term in the rule, and the topic is the rule consequent, in function block 16. Once the preceding steps have been accomplished, all the necessary data is at hand to finish setting up the natural language interface in function block 17. Setting up the dialog manager is accomplished according to the process described in copending patent application Serial No. 09/570,788.
